Comparative Study of Probabilistic Causal Effect Estimation Using Different Global Models

Master's Thesis

for acquiring the degree of Master of Science (M.Sc.)

in Statistics
at the School of Business and Economics
of Humboldt-Universität zu Berlin

submitted by

Dingyi Lai
Student Number: 615865

First Examiner: Prof. Dr. Stefan Lessmann

Second Examiner: Prof. Dr. Sonja Greven

Berlin, March 18, 2024

Comparative Study of Probabilistic Causal Effect Estimation Using Different Global Models

Candidate: Dingyi Lai[a,b], First Supervisor: Prof. Dr. Stefan Lessmann[b], Second Supervisor: Prof. Dr. Sonja Greven[b]

[a] *Masters of Science in Statistics, Humboldt-Universität zu Berlin, Berlin, 10178, Berlin, Germany*
[b] *School of Business and Economics, Humboldt-Universität zu Berlin, Berlin, 10178, Berlin, Germany*

Abstract

The estimation of causal effects becomes increasingly complex when interventions exert diverse influences across quantiles. Addressing this challenge, we introduce a novel global framework that seamlessly integrates causal analysis with prediction algorithms. Despite remaining theoretical gaps, we propose a standardized approach to answer this research question. This involves defining causal mechanisms via directed acyclic graphs, elucidating theoretical assumptions, conducting placebo tests to identify causal effects, and estimating probabilistic causal effects within the global causal framework. Through comparative analysis utilizing synthetic and real-world datasets, we demonstrate the potential of this framework to estimate varying causal effects across quantiles over time. While promising, ongoing refinement is necessary to enhance the framework's consistency and robustness.

Keywords: Causal Inference, Time-varying Effect, Probabilistic Prediction

1. Introduction

Causal inference aims to deduce, from a quantitative study, how specific outcomes in a particular scenario might have differed had a particular event of interest not occurred. In the context of a standard prediction formula, including an indicative variable representing the presence of the event, also interpreted as a treatment in a causal language, would theoretically serve a similar purpose. However, the application of various prediction algorithms within the realm of causal inference is nontrivial; it necessitates a comprehensive definition of a causal problem. Without this, it becomes challenging to persuade a decision-maker that a given policy or campaign has a causal effect on the targeted outcomes, as correlation does not inherently imply causality.

Philosophically speaking, the connection between causal inference and traditional prediction based on correlation lies in the existence of the concept of "causality". Practitioners cannot fully rely on the direct usage of deep learning prediction algorithms due to their "black-box" nature. As one of the ways out, the study of causal inference complements their unexplainability and lack of controllability or reliability.

Historically, in the 1920s, biologist Sewall Wright attempted to mathematically formulate causal relationships by integrating directed graphs and linear statistical models into a unified representation known as path analysis [60, 47]. Since then, three primary theoretical foundations developed by Lewis, Rubin, and Pearl have emerged. Lewis's Theory of Counterfactual Conditionals (LTC) is expounded from a philosophical perspective [37, 40, 38, 39], while Rubin's Causal Model (RCM), also known as the Neyman-Rubin causal model or Potential Outcome Framework (POF), is constructed starting from randomized controlled trials (RCT) [10, 36]. On the other hand, Pearl's Structural Causal Model (SCM) is associated with the work on directed acyclic graphs (DAG) [47, 49]. Although RCM and SCM both estimate causal effects, the question of their equivalence remains vague [43, 59, 48, 4]. One argument suggests that theorems provable in one system remain provable in the other, but the application of DAGs and the 4-step process of structural methodology (define, assume, identify, and estimate) adds marginal value in some cases [35, 33]. However, others argue that SCM primarily focuses on identifying "toy models" rather than estimation and inference, leading to oversimplification and detachment from reality [36].

In this thesis, a precise definition of a specific causal mechanism is provided through mathematical formulas and Directed Acyclic Graphs (DAGs) aligning with both Rubin's Causal Model (RCM) and Pearl's Structural Causal Model (SCM). The research question is: How would the treated units vary over time if they had not undergone treatment? In other words, what is the causal effect for each predicted horizon if the treatment influences the system after a specific time point? The research framework can, for instance, determine the impact of a new medication on patients' recovery rates over various time periods, examine the influence of environmental regulations on air pollution levels in the long run, or evaluate the effects of a promotional strategy on product sales over different time horizons. Although continuous A/B testing would fulfill a similar role, it is often costly and challenging to implement in some observational data in economics, health, or policy due to its unavailability.

To streamline the discussion, we adopt terminologies from RCM as the default. Simultaneously, we utilize DAGs from SCM, elucidating the transition from Rubin's language to Pearl's language. In Section 2, an overarching framework for causal analysis using RCM is delineated [58]. Given the thesis's emphasis on panel data, it is logical to commence with Difference-in-Difference (DiD) [12] and its extension, Synthetic Controls (SC) [3]. This method, also characterized by do-calculus proposed by Pearl [56, 62], sees recent extensions such as proximal SC [53]. Advancing one step further, a comprehensive integration of deep learning prediction and causal inference can be inferred, though not rigorously proven and directly employed in Grecov et al.'s work [29]. Given the fundamental problem of causal inference [32], addressing a causal problem necessitates testing the required assumptions through a placebo test. The extension from point to probabilistic estimates of potential outcomes is encompassed in DeepProbCP as well.

Four models, Causal Impact, TSMixer, DeepProbCP and Temporal Fusion Transformer (TFT), are introduced briefly in Section 3, with detailed descriptions of their architectures and the corresponding code utilized in the thesis. Among them, TSMixer and TFT, two state-of-the-art prediction algorithms, are employed within the same global causal paradigm as DeepProbCP, contrasting with the local causal framework represented by Causal Impact. Furthermore, three common error metrics, Mean Absolute Percentage Error (sMAPE), Mean Absolute Scaled Error (MASE) and Continuous Ranked Probability Score (CRPS), for evaluation are presented in this section. In Section 4, an empirical study is undertaken, employing both synthetic data and a real-world dataset. The panel data under discussion comprises multiple time series, with some designated as treated units and others as control units. When the structural data and the estimated results successfully pass the placebo test, the causal effect estimated by the predicted treated units and their observed values is considered reliable in theory. Consequently, it can be analyzed for decision-making in a more detailed manner.

2. Theoretical Background

The thesis delves into the potential correlation between causal language and predictive inference. Disregarding intricate and convoluted philosophical debates on the validity of ”causality” in nature, this chapter aims to elucidate the primary methodologies employed in causal inference when dealing with panel data. It also investigates the transition from estimating counterfactuals based on Synthetic Controls cross-sectionally to predicting counterfactuals over time using machine learning or deep learning. We assume that readers have a basic understanding of time series concepts, such as the distinctions between local and global univariate models, and between point and probabilistic forecasting.

The first subsection provides a brief introduction to the general framework of causal inference in Rubin's language, following the historical and genre-specific development outlined in the introduction section. Moreover, theoretical assumptions are delineated and extended from classical synthetic control models to an enhanced version discussed in Subsection 2.2, motivated by Shi et al.'s work [53]. Grecov et al.'s work [29] introduces a novel perspective on extending the application of generalized Synthetic Control from cross-sectional causal effect estimation to causal effects per treated unit and over time after an intervention, discussed in Subsection 2.3.

It is crucial to note that the estimated causal effect, whether by point estimation or per quantile, is deemed trustworthy when the data and the results pass the placebo test, under the specific causal relationship illustrated by Directed Acyclic Graphs (DAGs). If this framework is viable, a comparative study among different global forecasting methods can be conducted to enhance causal effect estimation in panel data.

2.1. Causal Inference Framework

The estimation of probability from past to future can be effectively managed through standard statistical analysis by inferring associations among variables, as long as experimental conditions remain constant. However, when conditions change, causal analysis becomes crucial for inferring probabilities. Pearl incorporates terminologies such as “error terms”, “stability”, “instrumental variables” and “explanation” into causal concepts. The distinction between causal and associational concepts lies in the fact that the former cannot be defined solely from the distribution alone [47]. Beyond merely answering “When did a change of this particular kind occur?” there is a continuous drive to understand “Why did it occur?” due to the ongoing pursuit of process robustness [57].

A potential outcome model within the RCM framework is widely adopted, starting with the assumption of Randomized Controlled Trials (RCTs) to define the Average Treatment Effect (ATE). However, as observational studies deviate from this ideal, the quality often relies on how well they approximate an RCT. To extend the RCT, unconfoundedness is introduced to eliminate any possible "back-door" effects for covariates $\mathbf{X}_i$. Considering propensity scores, this leads to building a naive estimator for ATE, either through matching or inverse-propensity weighting (IPW). Augmented IPW (AIPW), a more robust approach (double robustness), first estimates propensity scores non- or semi-parametrically and then estimates counterfactuals for treated and control units, computing the difference between the two predictions for each observation and averaging over these differences. The Cross-fitting modification of AIPW enables any $o_P(n^{-1/4})$-consistent machine learning method to be transformed into an efficient ATE estimator. Furthermore, when the ATE varies with the observed covariates $\mathbf{X}_i$, it becomes necessary to consider Conditional ATE (CATE) to estimate treatment heterogeneity [58].

For longitudinal data, panel data analysis methods are necessary, and in causal treatment effect estimation for time-varying data, there are primarily three scenarios: time-invariant treatment effect, time-varying treatment effect, and dynamic regimes. This master’s thesis will primarily focus on the second scenario. When a treatment occurs starting from a specific time point, such as T_0 in the next subsection, its effect is considered time-varying afterwards. Considering the simplest sample model, which assumes a constant treatment effect without treatment heterogeneity or dynamics, a two-way model and weighted two-way regression are commonly used. Difference-in-differences (DiD) is a classical tool among econometricians to capture causal effects based on the assumption of “common trends” or “parallel trends”. When parallel trends are no longer enforced, an approach called Synthetic Controls is proposed [3, 11, 58], which will be elaborated in the following section.

2.2. An Extension from (Proximal) Synthetic Control Models

Synthetic control modeling (SC) is widely used to predict counterfactual time series and estimate the causal effects of aggregated interventions [1]. This method is developed when a randomized controlled trial or A/B test is invalid due to unethical, expensive, or infeasible situations [7]. The traditional SC idea is to construct the artificial counterfactual of the treated unit in case it were not treated by estimating weights, such that a weighted average of the control units reconstructs the treated peers [13].

The standard SC model requires linearity and restricts the weights to be positive and add up to one [2]. Though it reduces the risk of overfitting, the assumed linearity is restrictive, and the focus on pre-treatment fit may not generalize out of sample [46].

The core challenges associated with SC lie within the realm of prediction. Therefore, approaches need to be developed specifically to accurately predict the counterfactual post-treatment outcome of the treated unit if it were absent from the treatment [46]. A modification by an elastic net estimator with flexible regularization, allowing weights to be negative and the sum is not required to be one [19]. While the assumption of linearity is theoretically justified in SC, a nonlinear and even non-parametric mechanism is proven to be valid and widely applied, especially in econometrics [52, 6, 5, 62]. If the focus is to capture the relationship between treated and control groups, lots of prediction models are capable of modeling the counterfactual prediction and estimating ATE with a thorough proof of asymptotic unbiasedness of machine learning models and consistency of the estimators for causal effect based on these models [46, 51, 20, 23].

Consider an observed set of time series with N units (e.g., countries, organizations, geographic areas, etc.) indexed by $i = 1, ..., N$. Let M index the treated units and let $|M|$ denote the number of treated units where $1 \leq |M| < N$. The rest of the time series is called control units and is indexed by C where the number of control units $1 \leq |C| < N$. Inside the group of control units, the units with similar pre-treatment outcome trajectories and covariates are called the "donor pool". Such a subset of control units is indexed by $D \subset C$ with $|D| < |C|$ being the number of donors, providing a good match to the treated units. Let T be the range of a time period indexed by $1, ..., T_0, ..., T$ with $T_0 < T$. Suppose that the treated unit $i \in M$ is not affected by the intervention of interest before T_0 but has been treated since $t > T_0$. If we denote the observed outcome of the treated units by $Y^{(1)}_{i \in M, t > T_0}$ and the corresponding counterfactual outcome by $Y^{(0)}_{i \in M, t > T_0}$, then our goal is to identify and estimate the average treatment effect on the treated unit (ATT) regarding the intervention of interest in period t as [1, 29, 52, 53]:

$$\tau_t = \mathbb{E}[Y^{(1)}_{i \in M, t > T_0} - Y^{(0)}_{i \in M, t > T_0} | M] \tag{1}$$

Since $Y^{(1)}_{i\in M,t>T_0}$ is known, the estimation process only requires determining $Y^{(0)}_{i\in M,t>T_0}$ conditioned on the treated units.

In the classical synthetic control methods, the consistency assumption [53] ensures that what we observe as the outcome is a specific manifestation of the potential outcome that would occur under the assigned treatment value.

Assumption 1 (Consistency). *For any unit i, in the pre-treatment period, treated units $Y_{i\in M,t\leq T_0} = Y^{(0)}_{i\in M,t\leq T_0}$ and control units $Y_{i\in C,t\leq T_0} = Y^{(0)}_{i\in C,t\leq T_0}$; in the post-treatment period, treated units $Y_{i\in M,t>T_0} = Y^{(1)}_{i\in M,t>T_0}$ and control units $Y_{i\in C,t>T_0} = Y^{(1)}_{i\in C,t>T_0}$ if treated.*

Furthermore, treated and control units are expected to adhere to the interactive fixed-effects model. The fundamental data-generating mechanism can be extended from a single treated unit to a group of treated units [1], where not only the treatment effect can be decomposed into a fixed and time-varying effect [53], but also covariates can be included [61]. When we try to capture the effect of unobserved factors which are illustrated as the upper blocks in Figure 1, we can either use a single interactive fixed effect $\mu_i^\top \lambda_t$ which would exclude δ_t from the blocks, or we can extend it to a more general form. The additive fixed effect $\delta_t + \zeta_i + \mu_i^\top \lambda_t$ in Equation 2 is also mentioned in [53], including both interactive fixed effect $\mu_i^\top \lambda_t$, time-varying common effect δ_t, and a fixed individual effect ζ_i.

Assumption 2 (Generalized fixed-effects model). *For unit i and time t:*

$$\begin{aligned} Y_{i\in M,t} &= \begin{cases} \delta_t + \zeta_{i\in M} + \mu_{i\in M}^\top \lambda_t + \xi_{i\in M}^\top X_{i\in M,t} + \epsilon_{i\in M,t}, & \text{if } t \leq T_0 \\ \beta_t + \delta_t + \zeta_{i\in M} + \mu_{i\in M}^\top \lambda_t + \xi_{i\in M}^\top X_{i\in M,t} + \epsilon_{i\in M,t}, & \text{if } t > T_0 \end{cases} \\ Y_{i\in C,t} &= \quad \delta_t + \zeta_{i\in C} + \mu_{i\in C}^\top \lambda_t + \xi_{i\in C}^\top X_{i\in C,t} + \epsilon_{i\in C,t}, \qquad \text{for all } t \end{aligned} \tag{2}$$

Where β_t is the time-varying treatment effect at each t caused since $t > T_0$, $\lambda_t \in \mathbb{R}^r$ is an $r \times 1$ vector of unmeasured common factors over time (stationary or non-stationary) and $\mu_i \in \mathbb{R}^r, i = 1, ..., N$ is an $r \times 1$ vector of unit-specific fixed factor loadings. Assuming that $X_{i,t}$ is an $r \times 1$ vector of observed associate external covariates per unit, $\xi_i^\top$ is an $r \times 1$ vector of associated unknown parameters. The error term $\epsilon_{i,t}$ satisfies [53]:

$$\mathbb{E}[\epsilon_{i,t}|\mu_i, \zeta_i, \delta_t, \lambda_t, X_{i,t}] = \mathbb{E}[\epsilon_{i,t}] = 0, \quad \text{for all } i \text{ and } t \tag{3}$$

Since one of the outstanding advantages of Pearl's Structural Causal Models (SCM) is an emphasis on a clear illustration of causal relationships among entities via DAG, the corresponding DAG is given in Figure 1. It effectively depicts the relationships among the treatment status I_t, the observed outcomes of the treated unit and control unit $Y_{i\in M,t}, Y_{i\in C,t}$, the unobserved common factor λ_t and its potential additive factor δ_t, observed covariates X_t before T_0, at T_0 and after T_0,

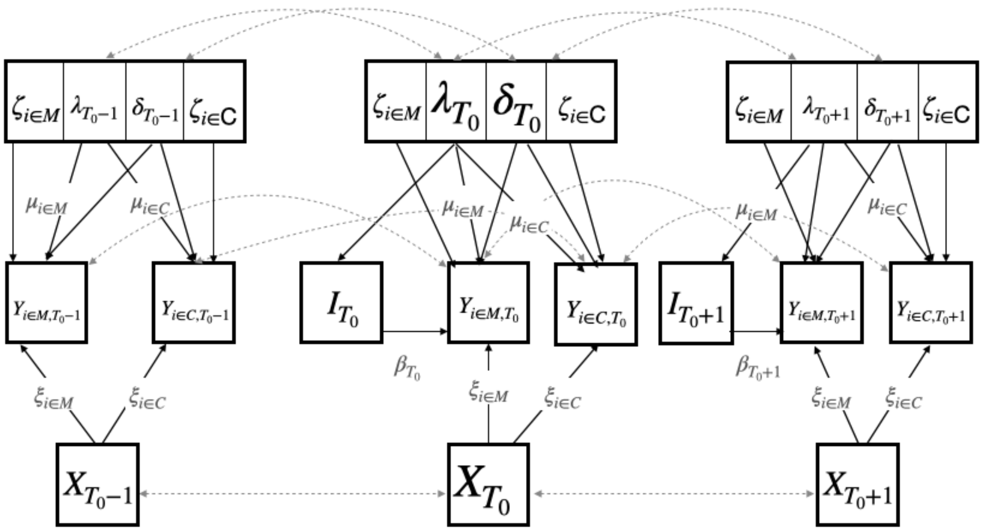

Figure 1: DAG for Synthetic Control Method

Deducing from Assumptions 1 and 2, the treatment effect can be calculated $\tau_t = Y^{(1)}_{i\in M,t>T_0} - Y^{(0)}_{i\in M,t>T_0} = \beta_t$, as for any i and t:

$$\begin{aligned} Y_{i\in M,t} &= \begin{cases} Y^{(0)}_{i\in M,t} = \delta_t + \zeta_{i\in M} + \mu^{\top}_{i\in M}\lambda_t + \xi^{\top}_{i\in M}X_{i\in M,t} + \epsilon_{i\in M,t}, & \text{if } t \leq T_0 \\ Y^{(1)}_{i\in M,t} = Y^{(0)}_{i\in M,t} + \beta_t, & \text{if } T_0 < t \leq T \end{cases} \\ Y_{i\in C,t} &= Y^{(1)}_{i\in C,t} = Y^{(0)}_{i\in C,t} = \delta_t + \zeta_{i\in C} + \mu^{\top}_{i\in C}\lambda_t + \xi^{\top}_{i\in C}X_{i\in C,t} + \epsilon_{i\in C,t}, \text{ for all } t \end{aligned} \tag{4}$$

One crucial assumption underlying the suggested SC methods is that the unobserved confounding impact on the treated unit can be balanced by a weighted average of the unobserved confounding impact on a specific group of control units, known as the "donor pool". However, not all control units should be selected into the "donor pool". Considering a linear relationship between $\mu^{\top}_{i\in M}$ and $\mu^{\top}_{i\in C}$, if D denotes units in the "donor pool", with $|D|$

being the number of donor units, we can assume $\alpha_i \in \mathbb{R}^d$ is a unit-specific d-dimensional vector. Accordingly, $\alpha_{i,i\in D}$ denotes a $d \times |D|$ matrix for $i \in D$ that are donor units. Such a key assumption can be formulated as:

Assumption 3 (Existence of synthetic control from classical synthetic control methods). *There exists a set of weights $\alpha_{i,i\in D}$ and $\beta_{i,i\in D}$such that*

$$\begin{aligned} \mu_{j\in M} &= \sum_{i\in D} \alpha_i \mu_i \\ \xi_{j\in M} &= \sum_{i\in D} \beta_i \xi_i \end{aligned} \tag{5}$$

Under Assumptions 1-3, we know $\mathbb{E}[Y^{(0)}_{j\in M,t}] = \mathbb{E}[\sum_{i\in D}(\alpha_i Y^{(0)}_{i,t} + \beta_i X_{i,t})]$. So the average treatment effect on the treated unit ATT in the post-treatment period can be formulated as [52]:

Proposition 1 *for any $t > T_0$*

$$\tau_t = \mathbb{E}[Y_{j\in M,t} - (\sum_{i\in D}(\alpha_i Y_{i,t} + \beta_i X_{i,t}))|\mu_i, \zeta_i, \delta_t, \lambda_t] \tag{6}$$

When it is extended to a non-parametric form as in the proximal causal inference framework [53], Assumptions 2 and 3 can be altered to:

Assumption 2' (No interference). $Y^{(0)}_{i,t} = Y^{(1)}_{i,t}$ *for any $i \in C$ and t*

Assumption 2' makes sure that there is no interference that affects the control groups at any time. This is the theoretical ground for a placebo test afterwards.

Assumption 3' (Existence of confounding bridge). *There exists a function $h(Y_{i\in D,t}, X_t)$ such that for any $t > T_0$:*

$$\mathbb{E}[Y^{(0)}_{j\in M,t}|\mu_i, \zeta_i, \delta_t, \lambda_t] = \mathbb{E}[h(Y^{(0)}_{i\in D,t}, X_t)|\mu_i, \zeta_i, \delta_t, \lambda_t] \tag{7}$$

where X_t are observed covariates for all units, since $Y^{(0)}_{j\in M,t}$ can also be partly explained by its covariates through $X_{j\in M,t}$.

Note that the function $h(Y_{i\in D,t}, X_t)$ in Assumption 3' is the non-parametric version of the synthetic control $\sum_{i\in D}(\alpha_i Y_{i,t} + \beta_i X_{i,t})$ where α_i and β_i is the parametric weight in Assumption 3, combining explainable parts from its own covariates $X_{j\in M,t}$. In other words, Assumption 3 can be seen as a special case of Assumption 3'.

Directly determining $h(\cdot)$ from Equation 7 is not feasible due to the unobservability of δ_t, λ_t. Therefore, there are mainly two ways to address this issue: either looking for additional proxy variables or assuming *Unconfoundedness* as always [53, 45, 55]. The *Unconfoundedness* as a classical identification assumption in causal inference indicates the ignorability of unobserved factors δ_t, λ_t. Because the identification of proxy variables is not easy, we assume *Unconfoundedness* in this thesis. But the alternative ways of using proxies should be kept in mind.

Besides the concerns on the unobserved factors δ_t, λ_t, there are other potential problems in such generalized synthetic control methods. For instance, if other time-confounding events, such as state-specific events occurring in the control units (e.g., states), happen at T_0, then the use of the synthetic controls method might be problematic. In such cases, reverting to the case-selection approach is preferred. However, if country-wide events affect all comparison control units (e.g., states) at T_0, the use of the synthetic controls method could be beneficial [18].

Simply put, under unconfoundedness, Assumptions 2' and 3', the following equation holds in the pre-treatment period:

$$\mathbb{E}[Y_{j\in M,t}] = \mathbb{E}[h(Y_{i\in D,t}, X_t)], \quad \forall t \leq T_0 \tag{8}$$

Therefore, in the pre-treatment period $t \leq T_0$, the confounding bridge function $h(\cdot)$ under Assumptions 1 and 3' satisfies the moment condition:

Proposition 2 (Moment condition for $h(\cdot)$) *for any* $t \leq T_0$

$$\mathbb{E}[Y_{j\in M,t} - h(Y_{i\in D,t}, X_t)] = 0 \tag{9}$$

Similarly, Proposition 1 can be transformed into:

Proposition 3 (Non-parametric identification of ATT) *for any* $t > T_0$

$$\tau_t = \mathbb{E}[Y_{j\in M,t} - h(Y_{i\in D,t}, X_t)] \tag{10}$$

Note that Proposition 3 tries to find a confounding bridge function $h(\cdot)$ given corresponding assumptions and observed covariates X_t to simulate $Y_{j\in M,t}$ on average, while Proposition 2 attempts to utilize such a confounding bridge function $h(\cdot)$ that is trained upon all information related to donor units $Y_{i\in D,t}$ and observed covariates X_t.

Furthermore, the definition above is grounded in RCM, aka. the potential outcomes framework. Recently, the theoretical gap between SCM and RCM has been bridged in this case [62], allowing the transformation of Equation 1 into Equation 11, accompanied by an illustrative Figure 1 [1, 29]:

$$\tau_t = \mathbb{E}[Y_{i\in M,t}|do(I_t = 1), I_t = 1] - \mathbb{E}[Y_{i\in M,t}|do(I_t = 0), I_t = 1] \tag{11}$$

where $I_t = 1$ denotes an intervention, and $I_t = 0$ implies no intervention at any t. The theorems from this paper [62] prove that such an $h(Y_{i \in D,t}, X_t)$ in Equation 10 does exist, ensuring:

$$\mathbb{E}[Y_{i \in M,t} | do(I_t = 0), I_t = 1] = \mathbb{E}[h(Y_{i \in D,t}, X_t) | Z_t] \tag{12}$$

Therefore, there is no debated discrepancy in theory between RCM and SCM under the above DAGs and assumptions. Furthermore, Directed Acyclic Graphs (DAGs) play a crucial role in this causal framework, aiding in the identification of key confounding events and facilitating the analysis of causal relationships among specific entities [18].

2.3. Global Model, Placebo Tests and Probabilistic Forecasting

The novel idea of DeepProbCP [29] is to map $h(Y_{i \in D,t}, X_t)$ in Equation 10 to $f(Y_{t \le T_0}, X_{t \le T_0})$ in Equation 13. Since Proposition 3 tries to capture information from the treated and donor units in the pre-treatment period, it is intuitive to use the pre-treatment data to simulate the counterfactual $Y^{(0)}_{i \in M, t > T_0}$.

Modification 1 (Global identification of the counterfactuals)

$$Y^{(0)}_{i,t > T_0} = f(Y_{t \le T_0}, X_{t \le T_0}) + u_t \tag{13}$$

From the dimension of time t, Equation 8 can be viewed as a regression model, since for each t, treated unit $j \in M$ can be constructed via control units and covariates; while Modification 1 extends it to a global univariate model, forecasting each unit based on historical treated and control units. Recall that when the target of prediction only involves one variable, it is called a univariate model; otherwise, i.e. there are more variables to forecast, it is called a multivariate model. The local univariate model trains each time series independently, and multidimensional covariates $x_{i,t}$ can be used for each of the N models. By contrast, the global univariate model estimates shared structure and parameters among all the time series. Such a global modelling strategy [9] trains the forecast model via all data before the intervention and then uses placebo tests to check whether the control units remain unchanged after the intervention.

Hence, there is a critical difference between the generalized synthetic control method and DeepProbCP: the former builds the non-parametric estimator per time point by comparing the two groups (vertically), but the latter constructs the predictor based on the pre-treatment data (horizontally). This discrepancy echoes the different definitions for evaluative designs on observational study in [18].

The estimation goal in this case is to forecast post-treatment time series based on the pre-treatment data in a normal predictive way. The modification will be in the application and interpretation. Theoretically, only when the predicted results pass the placebo test, we can trust the predicted treated units after intervention as the counterfactuals in causal analysis. To train such a predictor f, window-based training based on training data and validation data in the pre-treatment period is proposed in DeepProbCP [29]. Accordingly, ATT can be formulated as:

Modification 2 (Global identification of the ATT)

$$\tau_t = \mathbb{E}[Y_{j \in M,t} - f(Y_{t \leq T_0}, X_{t \leq T_0})], \quad \forall t > T_0 \tag{14}$$

The placebo test mentioned above is to confirm the null intervention effect over control units and the significant intervention effect over treated units, which is the preliminary condition to do the causal analysis based on such a forecaster [29, 15].

If the placebo test is significant, the ATT estimated by Equation 14 is more trustworthy. Generally, the greater the ratio of the prediction error of treated units after intervention by that of control units, the ATT is more significant.

The counterfactual $Y^{(0)}_{i,t+h}$ with a given horizon h is constructed via various variables, making it essentially a variable. Statistically, estimating Average Treatment Effects (ATT) through point forecasting and interval forecasting is prevalent. These methods assume a prior distribution or density, while probabilistic forecasting determines the distribution or density by fitting [25]. Probabilistic forecasts, being more realistic due to looser assumptions, not only provide an estimation of uncertainty essential for decision-makers but are also the focus of the M5 uncertainty competition [42]. Thus, to quantify the uncertainty of the prediction, quantile forecasting is also introduced in DeepProbCP compared to its precursor, DeepCPNet, although it may be too rough to distinguish between aleatoric and epistemic uncertainty [22]. Such an introduction can offer a bonus when the intervention occurs only within certain quantile ranges of the target.

Based on that, TTE can be calculated per quantile by predicting $Y^{(0)}_{i,t>T_0}$ per quantile. Take h as the length of the predicting horizon, Q as the number of quantiles, and N as the number of time series as always. DeepProbCP predicts various predetermined quantiles for each time series i, entailing training individual global LSTM models for each quantile q:

$$\hat{Y}^{(q)}_{i,t+1}, ..., \hat{Y}^{(q)}_{i,t+h} = f_q(Y_{t \leq T_0}, X_{t \leq T_0}) \tag{15}$$

Such quantile forecasts can be linked via spline interpolation to model the predicted counterfactual distribution. The spline functions can enforce smoothness through knots that link diverse polynomial functions:

$$s(z; a, b, c) = a + \sum_{l=0}^{L} b_l (y - c_l)^k_+ \tag{16}$$

where a is an intercept term, b is a vector of weights representing the slopes, k is the degree of the polynomial function indicating smoothness, c is a vector of knot positions, L denotes the number of pieces belonging to the hyperparameters of the spline. Running individual spline regressions with $k = 3$ (cubic) after quantile forecasting is an independent step.

DeepProbCP, as the start of such a framework, is constructed to perform probabilistic causal effect estimation by quantile forecasting. Similar to what M5 ”uncertainty” competition required, the prediction of different quantiles is sufficient to illustrate the distribution of future trajectories. To evaluate the performance of quantile forecasting, Quantile Loss (QL) serves as the loss function which is also called pinball loss. The intuition of the QL is to exert an asymmetrical penalty when the forecasts and actuals differ:

$$L_q(y, \hat{y}) = q(y - \hat{y})_+ + (1 - q)(\hat{y} - y)_+ \tag{17}$$

where q is the quantile that needs to be assessed, y and $\hat{y}$ are observed value and quantile prediction respectively. The rectified linear unit (ReLU) function $(\cdot)_+$ denotes $max(0, \cdot)$ to ensure the positivity of QL. Note that if q is set to 0.5, QL becomes the Mean Absolute Error (MAE). Minimizing MAE leads to forecasting the median of the forecast distribution.

In this thesis, we conduct a comparative study on different causal effect estimation methods with different global models. In other words, f_q is not constrained to neural-network-based methods like LSTM, but it can be extended to transformer-based models for probabilistic causal effect estimation under the novel framework.

Before we jump into the application of different models in causal effect estimation, we need to reclaim that this thesis is limited in many aspects.

First, though we know there is a possible way out through finding proxy variables to exclude the effects from unobserved confounding, a thorough review of how to do so across different literature and their underlying theory is awaited to be done. Second, to quantify the uncertainty of ATT, there should be a more detailed review of two inherently different sources of uncertainty - aleatoric and epistemic uncertainty [22]. Finally, we can extend the time-varying treatment effects to dynamic regime [30].

3. Methodology

On the one hand, to compare different global models for causal effect estimation, a benchmark is set using a Bayesian structural time series technique developed by Brodersen et al. [11]. This algorithm serves as a local and parametric alternative to DeepProbCP, lacking probabilistic estimation, but outperforms the artificial counterfactual (ArCo) using LASSO [29]. On the other hand, to compare different global predictive models for causal effect estimation, two benchmarks are compared. The first one is TSMixer, recently developed based on stacking multi-layer perceptrons (MLPs) to extract information efficiently by mixing operations along both the time and feature dimensions [14]. Another one is Temporal Fusion Transformers (TFT), a prevalent attention-based architecture for multi-horizon time series forecasting that combines high-performance forecasting with interpretable insights into temporal dynamics [41]. The implementation codes are available in $https://github.com/Dingyi-Lai/master\_thesis$.

3.1. Local Model: Causal Impact

Causal Impact, a local structural causal model, provides the flexibility to select from a wide array of potential controls by implementing a spike-and-slab prior on the regression coefficients set and allowing the model to incorporate an average over these controls. Subsequently, it calculates the posterior distribution of the counterfactual time series based on the target series' pre-intervention period values, in conjunction with the controls' values during the post-intervention period. The difference between the observed and predicted responses during the post-intervention period yields a semi-parametric Bayesian posterior distribution for the causal effect.

This model is implemented through the CausalImpact Python module [50], where the pre-treatment period and post-treatment period should be set properly.

3.2. Global Model

Unlike local models, global models share parameters learned across treated and control units without specifying the assumed data structure but fit a learning model to predict treated and control units separately.

All global models adopt the moving window transformation strategy, enabling modular training of the model. Figure 2 illustrates this approach using a real-world dataset comprising 62 series. In the off-the-shelf Python modules for TSMixer and TFT, moving window transformation strategy is inbuilt. For DeepProbCP, we use the *TFRecordReader* to restructure the dataset manually referred to the codes for DeepCPNet in *https : //bit.ly/3mWFbEO* [28]. Since each window is split into input and output windows, the last output window for each training time series is reserved as the validation set. It means that fixed-size windows with different starting points are constructed.

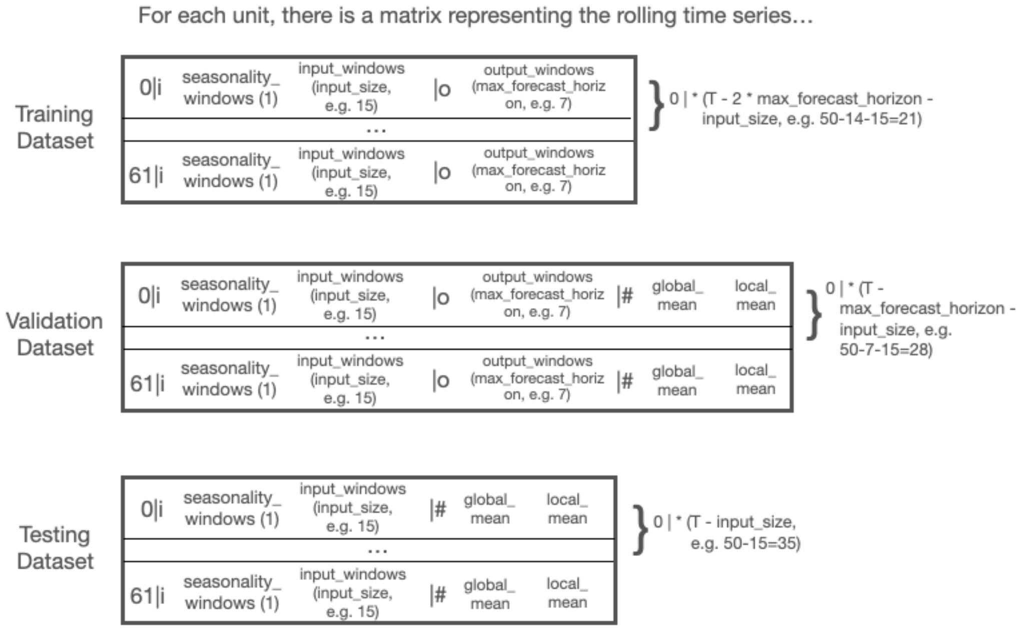
For each unit, there is a matrix representing the rolling time series...
Training Dataset
0|i
seasonality_ windows (1)
input_windows (input_size, e.g. 15)
|o
output_windows (max_forecast_horizon, e.g. 7)
...
61|i
0 | * (T - 2 * max_forecast_horizon - input_size, e.g. 50-14-15=21)
Validation Dataset
|#
global_ mean
local_ mean
0 | * (T - max_forecast_horizon - input_size, e.g. 50-7-15=28)
Testing Dataset
0 | * (T - input_size, e.g. 50-15=35)

Figure 2: Dataset Structure for 911 Emergency Call Dataset

3.2.1. MLPs-based: TSMixer

TSMixer is a comprehensive framework designed for multivariate time series forecasting, leveraging modular architecture and advanced techniques to achieve state-of-the-art performance across diverse forecasting tasks.

Key components include the Time-mixing MLP, which focuses on modeling temporal patterns within the data using a fully-connected layer, activation function, and dropout mechanism. Additionally, the Feature-mixing MLP operates across time steps to leverage covariate information, enhancing the model's ability to capture complex relationships between variables. To adapt to various forecasting horizons, TSMixer incorporates Temporal Projection, mapping input time series to the desired forecast length. Residual Connections between Time-mixing and Feature-mixing layers enhance the model's capacity to capture long-term dependencies and mitigate the vanishing gradient problem. For efficient training and optimization, TSMixer implements 2D normalization on both time and feature dimensions. This technique stabilizes training and ensures effective learning, even in the presence of complex temporal and covariate structures.

The reason for choosing it is its purported superior performance compared to transformer-based models in pure prediction tasks [14]. It is developed by Google Research [27] and is utilized for prediction within the global causal framework presented in the thesis.

3.2.2. LSTM-based: DeepProbCP

In the training process, DeepProbCP also employs the moving window transformation strategy. The Quantile Loss (QL), as formulated in Equation 17, is computed per window, and the loss is minimized using the Continuous Coin Betting (COCOB) optimizer [31]. Unlike Adam or Adagrad optimizers, COCOB does not require a predetermined learning rate hyperparameter, as it can self-tune its learning rate, leading to faster convergence. Hyperparameter tuning is conducted using the Sequential Model-based Configuration (SMAC) algorithm, a generalization of Sequential Model-based Optimization (SMBO), which iterates between model fitting and configuration selection and has demonstrated superior performance in empirical studies [34].

To handle the seasonality and trend of time series, DeepProbCP adopts the Seasonal Exogenous (SE) approach. This method extracts seasonal components more effectively, especially in homogeneous time series, compared to deseasonalized approaches. It utilizes the output of the pre-processing layer along with seasonal components as externals in each input window [8].

The implementation of DeepProbCP is inherited from DeepCPNet [28] and is updated in its optimization criterion, transitioning from regularized Mean Absolute Error (MAE) to Continuous Ranked Probability Score (CRPS).

3.2.3. Transformer-based: Temporal Fusion Transformers (TFT)

The Temporal Fusion Transformer (TFT) aims to tackle the issue of forecasting variables-of-interest across multiple future time steps, a task often characterized by a complex amalgamation of inputs without prior knowledge of their interactions with the target variable. It leverages recurrent layers for localized processing and interpretable self-attention layers for capturing long-term dependencies. Additionally, it incorporates specialized components designed to identify pertinent features and suppress irrelevant ones.

The algorithm is purported to surpass DeepAR and MQRNN, two other popular global probabilistic forecasters [41]. TFT is also developed by Google Research [26]. However, we utilize an off-the-shelf deep learning framework called *PyTorch Lightning* to implement the model, with hyperparameter tuning performed using the *Optuna* tuning framework.

3.3. Evaluation Metrics

In the empirical study detailed in Section 4, we generate a set of synthetic data and utilize a real-world dataset. While only the actual counterfactual treated units in synthetic data can be known beforehand, we can still apply the same metrics for evaluation. The primary distinction arises with real-world data, where we can only compute the errors for the control units.

For the synthetic data, given our prior knowledge of the actual treated units in the absence of intervention, we assess predictive performance using the true treated counterfactual, observed control units, predicted treated units, and predicted control units. To measure accuracy, we employ the Symmetric Mean Absolute Percentage Error (sMAPE) and the Mean Absolute Scaled Error (MASE), defined as follows [28]:

$$\text{sMAPE} = \frac{2}{h} \sum_{t=T_0+1}^{T_0+h} \frac{|\hat{Y}_t^{(0)} - Y_t^{(1)}|}{(|Y_t^{(1)}| + |\hat{Y}_t^{(0)}|)} \tag{18}$$

$$\text{MASE} = \frac{1}{h} \frac{\sum_{t=T_0+1}^{T_0+h} |\hat{Y}_t^{(0)} - Y_t^{(1)}|}{\frac{1}{T_0-S} \sum_{t=S+1}^{T_0} |Y_t^{(1)} - Y_{t-S}^{(1)}|} \tag{19}$$

where T_0 is the length of training set, h is the forecasting horizon, S is the length of seasonality, $Y_t^{(1)}$ is the observed matrix of the targets at time t, and $\hat{Y}_t^{(0)}$ is the predicted counterfactual matrix of the targets at time t. These metrics provide measures of the accuracy of the predicted counterfactuals compared to the observed values, taking into account both the magnitude and direction of the errors.

For those models that can forecast per quantile, Continuous Ranked Probability Score (CRPS) can be a good option. CRPS provides a measure of the overall predictive accuracy of a probabilistic forecast, taking into account both the calibration and sharpness of the forecast distribution. A lower CRPS indicates better predictive performance. It can be defined as [29]:

$$\text{CRPS} = \int_0^1 2L_q(\widehat{\text{CDF}}^{-1}(q), y)\, dq \tag{20}$$

where q is the quantile level, $\widehat{\text{CDF}}^{-1}$ is the cumulative distribution function represented by the quantile function, y is the target observation at which the CRPS is evaluated, and L_q is the pinball loss in Equation 17. Notice that only DeepProbCP and TFT give probabilistic forecasts in the thesis.

Additionally, it should be noted that all metrics presented in the thesis are averaged for measurement.

4. Empirical Study

To illustrate the overall idea of estimating time-varying probabilistic causal effects, two types of data are utilized. For the synthetic dataset, we generate 24 datasets, ranging from short time series (length is 60) to long time series (length is 222), and from homogeneous intervention over 0.9 quantiles (adding one unit standard deviation of treated units before intervention) to heterogeneous intervention over 0.9 quantiles (adding a random number between 0.7 to 1.5, multiplied by one unit standard deviation of treated units before intervention). The dataset also spans from a few time series (amount is 10), medium time series (amount is 101) to many time series (amount is 500), and from a linear data generation process (autoregressive regression, i.e. AR) to nonlinear structure (self-exciting threshold autoregressive, i.e. SETAR) [31, 29, 28]. Since the data is generated based on unemployment rate data (`UNRATE` from the `DATA_USEconModel` dataset available in the MATLAB software [44, 29]), it is possible that the trend of treated units could be significantly different from that of control units due to the randomness. Hence, the placebo test should contribute to filtering out those unqualified prediction models that fail to capture the relationship between the treated and the control.

As for the real-world data, we choose to replicate the `911 Emergency Call Dataset` which is available on Kaggle [16] from Grecov et al.'s previous work [28]. The `911 Emergency Call Dataset` encompasses emergency calls related to EMS, traffic, and fire, detailed in 88 distinct types of codes. Covering 62 municipalities in Montgomery County, United States, the dataset spans from December 2015 to July 2020. To address data sparsity, we aggregate the original daily observations into monthly levels and categorize the 88 codes into EMS, traffic, and fire categories. This dataset is specifically employed to investigate the impact of COVID-19 lockdown measures on the demand for 911 emergency calls. The lockdown restrictions were implemented from January 2020, setting the post-intervention period from January 2020 to July 2020 and the pre-intervention period from December 2015 to December 2019. To align heuristically [28, 29], the training output window size for `911 Emergency Call Dataset` is set to 7, with the corresponding training input window size set to 15. By contrast, the output window size for synthetic data is 12 with the same training input window size set being 15. Both size of seasonality is set to 12 since they are both monthly data.

4.1. Synthetic Data

Prior to simulation, it is essential to conduct a thorough examination of the underlying dataset, which in this case is the unemployment rate data mentioned previously. For more details, refer to Appendix 6.1.

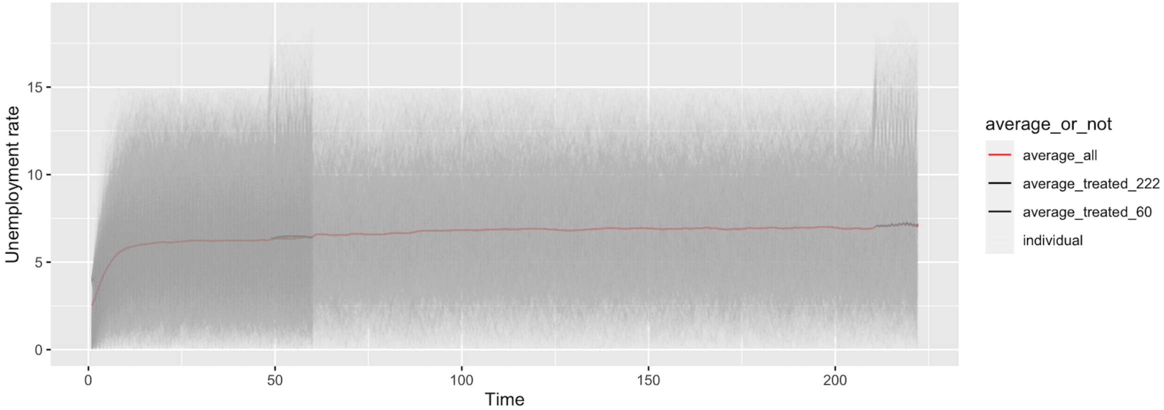

Figure 3: Simulated Time Series Based on Unemployment Rate Data

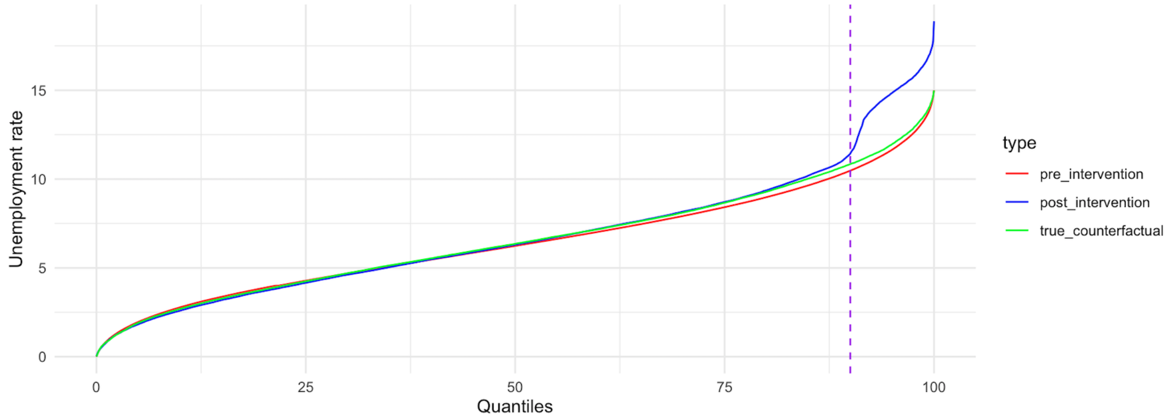

Figure 4: Quantile Distribution for the Simulated Treated Units

The simulated time series are depicted in Figure 3. Two notable anomalies can be observed in the trends, occurring from time = 48 to time = 60 and from time = 210 to time = 222. Given that both homogeneous and heterogeneous interventions are manually introduced for the treated units exceeding the 0.9 quantile threshold, a noticeable increase is evident in Figure 4.

The experimental results are presented in Tables 1, 2, and 3. Regarding point prediction accuracy, Causal Impact generally performs well when there are few time series, and the data generation process follows a linear model. However, when there are many units, TFT consistently outperforms other models, efficiently capturing nonlinear relationships over time. In contrast, DeepProbCP does not demonstrate a clear advantage in either point or probabilistic prediction compared to TFT. TSMixer, on the other hand, is claimed to effectively utilize persistent temporal patterns, cross-variate information, and auxiliary features in some cases compared to transformer-based models. However, the experimental results do not indicate its superiority in either prediction or the estimation of causal effects in the global framework. This might be attributed to TSMixer's ability to prevent overfitting, especially when the target is not correlated with other covariates, which is not the case in the global causal framework discussed in the thesis [14].

Table 1: Performance Metrics (sMAPE and MASE) for Different Models

Dataset					**sMAPE**				**MASE**			
DGP	Intervention	No. Series	Length	No. Data Points	Causal Impact	TSMixer	DeepProbCP	TFT	Causal Impact	TSMixer	DeepProbCP	TFT
Linear ARIMA	Homogeneous	10	60	600	0.3890	0.6155	**0.3054**	0.3410	1.2867	2.1078	**1.1181**	1.2608
			222	2220	0.3275	0.4159	0.4256	**0.2747**	1.3547	1.3947	1.7907	**1.2310**
		101	60	6060	0.3621	0.5555	**0.3532**	0.4012	**1.0618**	1.8794	1.3188	1.3344
			222	22422	0.3967	0.2581	0.2921	**0.2363**	1.6103	1.0535	1.2469	**0.9856**
		500	60	30000	**0.2464**	0.8207	0.3493	0.3626	**0.7630**	2.9686	1.1462	1.1601
			222	111000	0.2537	0.2899	0.2681	**0.1938**	0.9772	1.1348	1.1612	**0.7698**
	Heterogeneous	10	60	600	**0.2796**	0.5353	0.3407	0.3267	**0.8978**	1.7136	1.1802	1.0998
			222	2220	**0.2487**	1.5442	0.3058	0.2807	**0.8806**	0.6973	1.1470	0.9957
		101	60	6060	0.3868	0.7806	**0.3463**	0.4511	**1.1458**	2.7484	1.1975	1.4383
			222	22422	0.3815	0.2650	0.2689	**0.2324**	1.5442	1.0804	1.1084	**0.9451**
		500	60	30000	**0.2679**	0.7788	0.3511	0.4467	**0.8049**	2.7807	1.1162	1.3795
			222	111000	0.2551	0.2793	0.2769	**0.2030**	0.9366	1.0878	1.1759	**0.8027**
Non-Linear SETAR	Homogeneous	10	60	600	0.7256	1.1418	**0.5789**	0.6368	0.9243	2.4138	**0.8581**	1.1379
			222	2220	0.5195	0.6895	0.5183	**0.4588**	0.8869	1.1467	0.8840	**0.8023**
		101	60	6060	0.6259	1.0817	0.5944	**0.5358**	1.1452	2.5556	1.1702	**1.0899**
			222	22422	0.7214	0.4731	0.5512	**0.4068**	1.1190	0.8347	0.9926	**0.6795**
		500	60	30000	**0.3926**	1.1429	0.5341	0.6017	**0.7466**	3.2738	1.0699	1.1832
			222	111000	0.4724	0.4283	0.5145	**0.3528**	0.8723	0.8231	1.0062	**0.6559**
	Heterogeneous	10	60	600	0.6760	0.8757	0.5581	**0.5287**	1.2961	1.9791	1.1594	**1.1095**
			222	2220	0.6566	0.5876	0.4876	**0.3987**	1.4115	1.0855	1.0924	**0.7691**
		101	60	6060	0.5131	0.9086	0.5372	**0.4892**	**0.9914**	2.0762	1.0488	0.9917
			222	22422	0.8339	0.5499	0.5428	**0.4310**	1.6771	1.0207	1.0631	**0.7948**
		500	60	30000	**0.3690**	1.0963	0.5347	0.5165	**0.7466**	3.2089	1.0877	1.0765
			222	111000	0.4741	0.4702	0.5390	**0.3476**	0.8621	0.8759	1.0475	**0.6386**

Note: The bold font indicates the smallest error in the performance comparison.

In Table 2, not all the prediction models can pass the placebo Test, even though the intervention is actually added as a treatment during the generation process. The placebo test is implemented based on the codes provided by Grecov et al. [28]. However, an alternative method for conducting the test is mentioned in one of their recent papers [29], referring to Chernozhukov et al.'s work [15], but the codes are unavailable online. This is why we adhere to the traditional test used by Grecov et al. in [28]. Theoretically, the placebo test should help us judge whether a particular prediction model fits the causal relationship we want to test. If it does not pass, it suggests that either the prediction model is unsuitable, or the data fails to reflect the causal relationship. This could lead to another debate on the issue of ’research degrees of freedom’ [54]. Another possible reason could be the incapability of the placebo Test, which could be further improved in such a case [21].

Table 2: Performance Metrics (CRPS) for Different Models and Placebo Test

Dataset					**CRPS**		**Placebo Test**			
DGP	Intervention	No. Series	Length	No. Data Points	DeepProbCP	TFT	CausalImpact	TS-Mixer	DeepProbCP	TFT
Linear ARIMA	Homogeneous	10	60	600	0.1028	**0.0702**	0.4337	0.0640*	0.2328	0.5224
			222	2220	0.1225	**0.0644**	0.0005***	0.0013***	0.3394	0.7567
		101	60	6060	0.1094	**0.0951**	0.0000***	0.0159**	0.1252	0.0011***
			222	22422	0.0870	**0.0509**	0.0000***	0.0002***	0.0000***	0.2259
		500	60	30000	0.1023	**0.0778**	0.0000***	0.4238	0.0005***	0.0000***
			222	111000	0.0815	**0.0351**	0.8229	0.0146**	0.0267**	0.0036***
	Heterogeneous	10	60	600	0.1044	**0.0658**	0.0146**	0.2334	0.0000***	0.0005***
			222	2220	0.0887	**0.0497**	0.0026***	0.0161**	0.0031***	0.0338**
		101	60	6060	0.1047	**0.1002**	0.0063*	0.0093**	0.3868	0.0054*
			222	22422	0.0762	**0.0447**	0.0005***	0.0269**	0.1160	0.0000***
		500	60	30000	**0.0994**	0.1047	0.0000***	0.0146**	0.0034***	0.0000***
			222	111000	0.0838	**0.0368**	0.0000***	0.0339**	0.6423	0.0064***
Non-Linear SETAR	Homogeneous	10	60	600	**0.1681**	0.1870	0.1998	0.9097	0.0079**	0.0000***
			222	2220	0.1612	**0.1004**	0.0900*	0.2859	0.3804	0.4238
		101	60	6060	0.2049	**0.1276**	0.4697	0.0640*	0.0049***	0.0515*
			222	22422	0.1756	**0.0755**	0.0000***	0.4238	0.0034***	0.0341**
		500	60	30000	0.1704	**0.1460**	0.2307	1	0.6772	0.2151
			222	111000	0.1584	**0.0652**	0.1858	0.4358	0.0000***	0.0068**
	Heterogeneous	10	60	600	0.1519	**0.1026**	0.1679	0.5693	0.0016**	0.0351**
			222	2220	0.1582	**0.0721**	0.2334	0.0522*	0.1763	0.3394
		101	60	6060	0.1679	**0.1061**	0.9697	0.9097	0.4286	0.0000***
			222	22422	0.1653	**0.0873**	0.0000***	0.6038	0.4813	0.1313
		500	60	30000	0.1680	**0.1236**	0.2661	0.5693	0.0000***	0.3013
			222	111000	0.1692	**0.0638**	0.0006***	0.6221	0.0000***	0.0189**

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: The bold font indicates the smallest error in the performance comparison.

Different from the calculation of sMAPE for all treated and control units given the actual counterfactual treated units and observed control units in Table 1, the Average Treatment Effect on the Treated (ATT) is only estimated for those treated units over the 0.9 quantiles compared to the true counterfactual because this is how we introduce the intervention. In Table 3, it is difficult to determine which prediction model can capture the probabilistic causal effect more accurately and consistently.

Table 3: ATT (q >0.9) for the Treated and its Evaluation Based on sMAPE

Dataset					**ATT (q >0.9)**					**sMAPE for ATT**			
DGP	Intervention	No. Series	Length	No. Data Points	Actual	Causal Impact	TS-Mixer	DeepProbCP	TFT	Causal Impact	TS-Mixer	DeepProbCP	TFT
Linear ARIMA	Homogeneous	10	60	600	1.37	0.41	2.11*	−0.31	0.30	**0.43**	1.16	1.23	1.02
			222	2220	2.54	6.34***	2.35***	2.35	7.01	0.85	0.49	**0.21**	0.93
		101	60	6060	2.08	1.47***	1.05**	1.31	2.55***	0.35	0.76	0.59	**0.20**
			222	22422	2.59	1.24***	2.89***	2.58***	3.11	0.82	**0.13**	0.14	0.18
		500	60	30000	2.26	2.70***	−1.90	1.68***	2.46***	0.23	1.62	0.53	**0.15**
			222	111000	2.57	2.09	1.67**	1.58**	3.15***	0.20	0.55	0.64	**0.20**
	Heterogeneous	10	60	600	0.75	2.38**	3.3	−2.00***	2.79***	**1.02**	1.15	1.31	1.09
			222	2220	−0.15	−0.54***	1.53**	−0.70***	−0.81**	0.98	1.34	**0.97**	1.09
		101	60	6060	3.07	2.58*	−0.74**	3.74	3.70*	**0.17**	0.87	0.37	0.18
			222	22422	2.35	0.88***	2.42**	2.89	3.66***	1.01	0.28	**0.22**	0.42
		500	60	30000	1.57	2.13***	−2.61**	1.39***	2.15***	**0.30**	1.44	0.59	0.31
			222	111000	2.12	1.97***	3.04**	1.66	3.26***	**0.11**	0.34	0.59	0.42
Non-Linear SETAR	Homogeneous	10	60	600	2.49	1.93	−4.93	5.31**	2.90***	1.10	1.60	0.96	**0.81**
			222	2220	2.90	2.26*	−2.53	−0.03	1.27	**0.49**	1.82	1.59	0.80
		101	60	6060	3.04	0.75	−7.62*	0.48***	2.58*	1.29	1.91	1.36	**0.40**
			222	22422	2.88	2.10***	4.48	2.84***	4.92**	0.55	0.42	**0.23**	0.52
		500	60	30000	3.16	4.45	−13.01	2.03	2.86	0.33	2.0	0.51	**0.22**
			222	111000	3.05	2.47	4.93	3.19***	5.84**	0.21	0.47	**0.15**	0.59
	Heterogeneous	10	60	600	2.99	−1.96	−3.63	1.17**	1.56**	1.52	1.52	0.92	**0.78**
			222	2220	1.70	5.25	3.89*	1.87	5.47	1.19	1.28	**0.73**	1.27
		101	60	6060	3.06	0.63	−7.03	1.72	3.05***	1.25	1.86	0.82	**0.14**
			222	22422	3.99	−1.41***	2.90	3.77	6.06	1.79	0.43	0.41	**0.39**
		500	60	30000	2.18	3.31	−15.31	1.42***	2.35	0.40	2.0	0.49	**0.18**
			222	111000	2.72	2.27***	3.76	1.75***	5.42**	**0.19**	0.31	0.65	0.61

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$
Note: The bold font indicates the smallest error in the performance comparison.

Methodologically, we should select the forecast model that passes the placebo Test and provides the lowest sMAPE for the Average Treatment Effect on the Treated (ATT) estimation, as this indicates a higher level of trust in the estimated causal effect. The reason for the inconsistency could be either the prediction model fails to integrate well into the causal analysis framework, or the placebo test is flawed. Therefore, given all the tools we have discussed so far, we suggest estimating the probabilistic causal effect in a global model through the following steps:

1. Analyze the causal relationship qualitatively and draw a Directed Acyclic Graph (DAG) to illustrate it.
2. Predict the counterfactual treated units and calculate the sMAPE, MASE, and CRPS (if possible) to evaluate prediction performance.
3. Conduct a reliable placebo test to determine whether the estimated causal effect is significant using a particular forecasting model.
4. If the placebo test is passed, calculate the probabilistic causal effect and generate the corresponding sMAPE.
5. Select the probabilistic causal effect when the placebo test is passed and it provides the smallest sMAPE for TTE estimation.

4.2. Real-world Data

Figure 5 replicates the negative causal effect on the treated as observed in Grecov et al.'s work [28], confirming the accurate implementation of their proposed global probabilistic causal estimation framework represented by DeepProbCP. A possible explanation for the negative causal effect could be that COVID-19 lockdown measures restricted the demand for 911 emergency calls, possibly due to factors such as fatalities or people's reluctance to visit hospitals during that period.

Noticeably, when we utilize spline interpolation to connect all predicted quantiles, we encounter the well-known issue of *quantile crossing* [24], as detailed in Appendix 6.2. It could be improved by using the family of conditional quantile functions [24] in the future work.

Similarly, prediction performance can also be evaluated for the control units, given the absence of actual counterfactuals for the treated units in the real-world data. A comparison is provided alongside the results from the Placebo Test. In the real-world scenario, despite the dataset being limited in both length and size, DeepProbCP surpasses other models in point estimation, even with its training criterion based on the CRPS. However, TFT demonstrates comparable performance in point estimation to DeepProbCP while exhibiting a lower CRPS. Surprisingly, Causal Impact fails to pass the placebo test, while all the global models appear to identify a significant causal effect in the 911 emergency calls data.

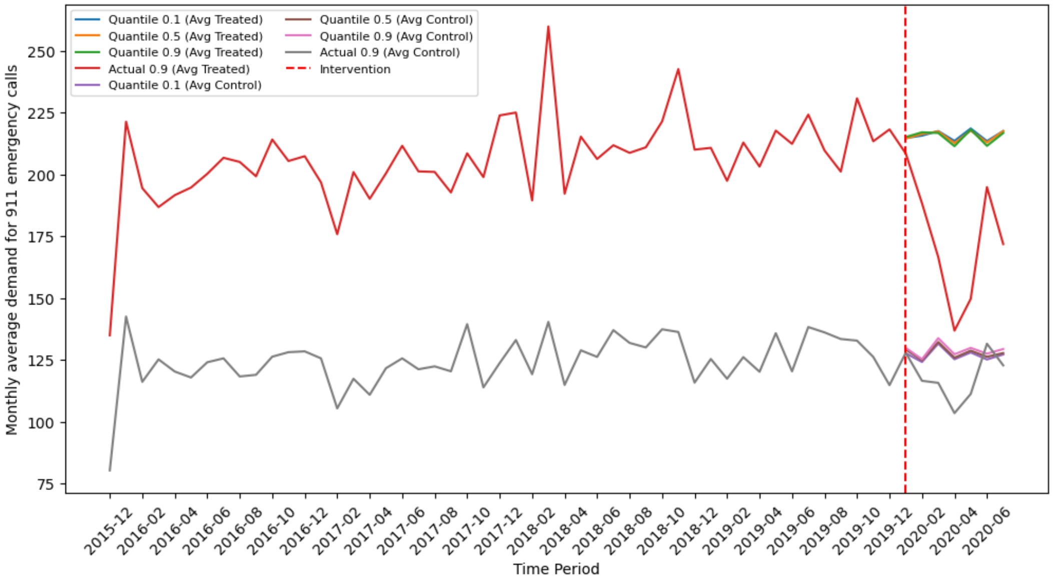

Figure 5: Counterfactual Results for Real-world Data from DeepProbCP

Table 4: Error Results Only for the Controls and the Placebo Test

Dataset: 911 emergency calls		Causal Impact	TSMixer	DeepProbCP	TFT
Prediction Performance on the Controls	**sMAPE**	0.4606	0.2574	**0.1929**	0.1950
	MASE	1.6628	1.1875	**0.8742**	0.8783
	CRPS	-	-	0.0423	**0.0262**
Placebo Test		0.2328	0.0017***	0.0040***	0.0053***

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: The bold font indicates the smallest error in the performance comparison.

Table 5 illustrates the causal effect for each quantile. Increasing the number of estimated quantiles would improve future analyses. Nonetheless, a discernible trend emerges, especially for the 0.9 quantile, revealing a significant negative causal effect on the demand for 911 emergency calls following COVID-19 lockdown measures, particularly in counties with initially higher demand.

Table 5: Estimation of the Average Treatment Effect on the Treated per Quantile

Quantile	Causal Impact	TSMixer	DeepProbCP	TFT
0.1	-6.5388	-56.3182	-32.7209	-54.2892
0.5	1.1404	-96.8799	-43.8783	-79.6912
0.9	-10.7043	-137.2120	-126.3976	-171.0350

The varying causal effects observed across quantiles would assist policy analysts or decision-makers in considering heterogeneous effects. Therefore, integrating probabilistic estimation into causal analysis holds significant potential.

5. Conclusion

This thesis aims to estimate the causal effect for each quantile at various predicted horizons, particularly when the treatment affects the system after a specific time point. It begins by reviewing the historical evolution of causal inference and selecting suitable scenarios and corresponding methodologies. Inspired by cutting-edge research, we seek to leverage the advantages offered by prediction algorithms to integrate them into the global causal framework.

The introduction of probabilistic estimation and placebo tests facilitates the fusion of causal analysis and prediction tasks. However, there remains a theoretical gap in mapping causal analysis for panel data to the novel causal framework enabled by prediction models. This gap may entail inconsistencies in results observed especially when synthetic data is employed in empirical studies.

Among the four models selected for comparison, Causal Impact demonstrates promising performance in cases involving short time series generated linearly but struggles to capture causal effects in more complex real-world scenarios. Conversely, while TSMixer initially appears less effective in simpler datasets, it surpasses Causal Impact in more complex data scenarios, offering superior performance.

DeepProbCP may hold its own advantages, extending the global causal framework from point estimation to probabilistic estimation. However, by adopting advanced prediction algorithms such as TFT while retaining the logic of probabilistic causal effect estimation, improved performance can be achieved, evident in passing placebo tests, TTE estimation, and evaluation based on CRPS.

There is ample room for future exploration and refinement in this area of study. For example, alternative global prediction models for time series, such as LightGBM, could be investigated to assess their efficacy in causal inference tasks. Furthermore, expanding the range of quantiles beyond the current selection of 0.1, 0.5, and 0.9 could provide a more comprehensive understanding of the quantile distribution for each forecasting horizon and make better use of spline interpolation.

Addressing the issue of quantile crossing could involve the implementation of conditional quantile functions, which may enhance the accuracy and stability of predictions. Similarly, improvements to the placebo test methodology could lead to more robust identification of causal effects.

It is essential to prioritize the development of mathematical frameworks to bridge existing theoretical gaps. Failure to do so may result in misspecification or oversight of true causal effects, exacerbated by the inherent flexibility in research methodologies. Additionally, in theoretical terms, indiscriminate inclusion of all control units in training datasets may introduce bias or distort the estimation of causal effects if certain units are unsuitable controls [17].

6. Appendix

6.1. Descriptive Analysis for Unemployment Data

To have a deep understanding of the unemployment data from MATLAB [44, 29], an autocorrelation analysis, including the examination of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), plays a crucial role in uncovering underlying patterns, identifying seasonality, and assessing the presence of randomness within the data.

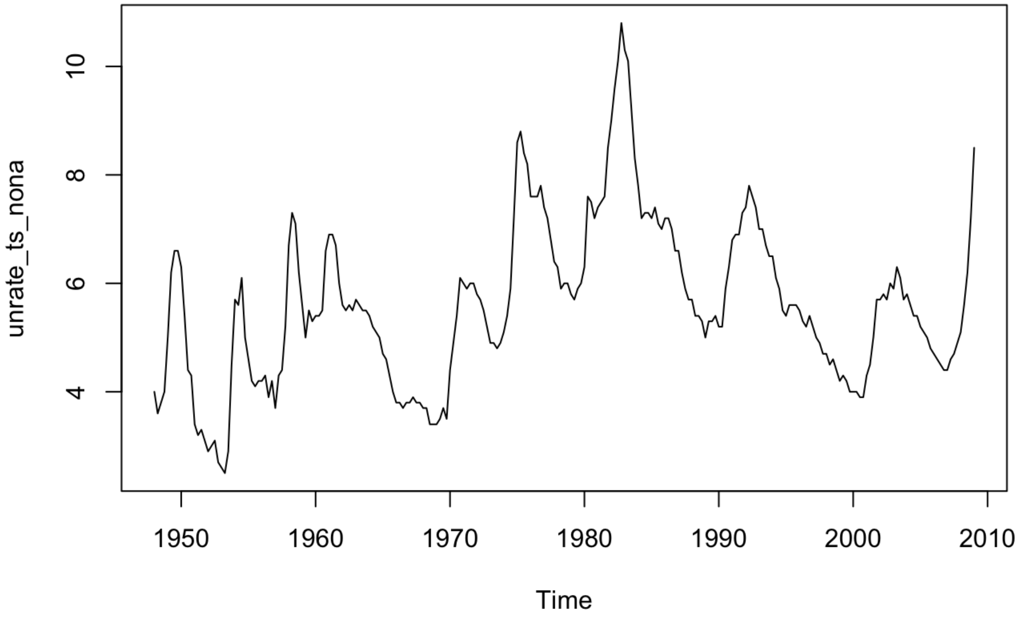

Figure 6: Line Plot for Unemployment Rate Data

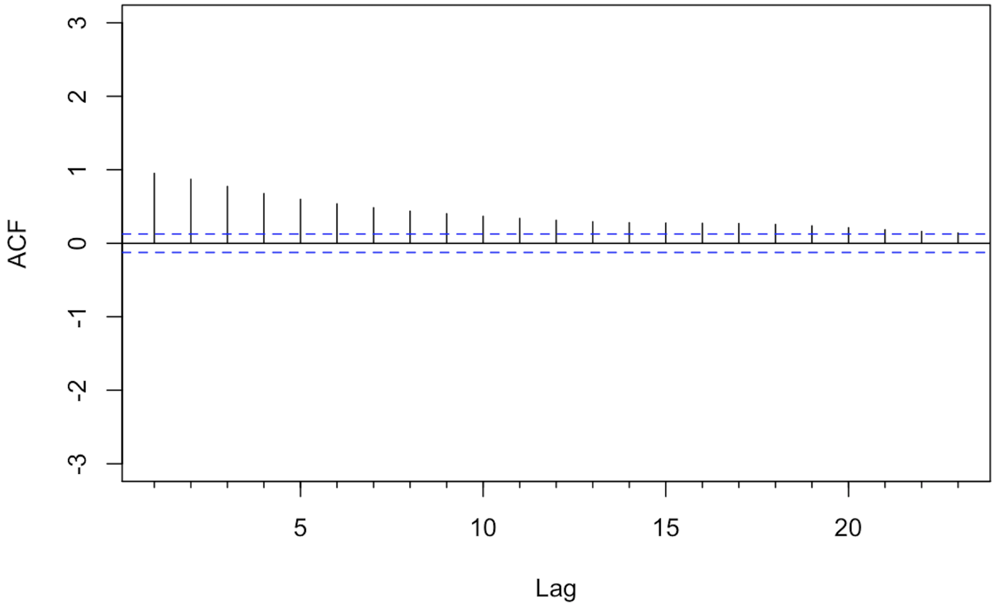

Figure 7: ACF for Unemployment Rate Data

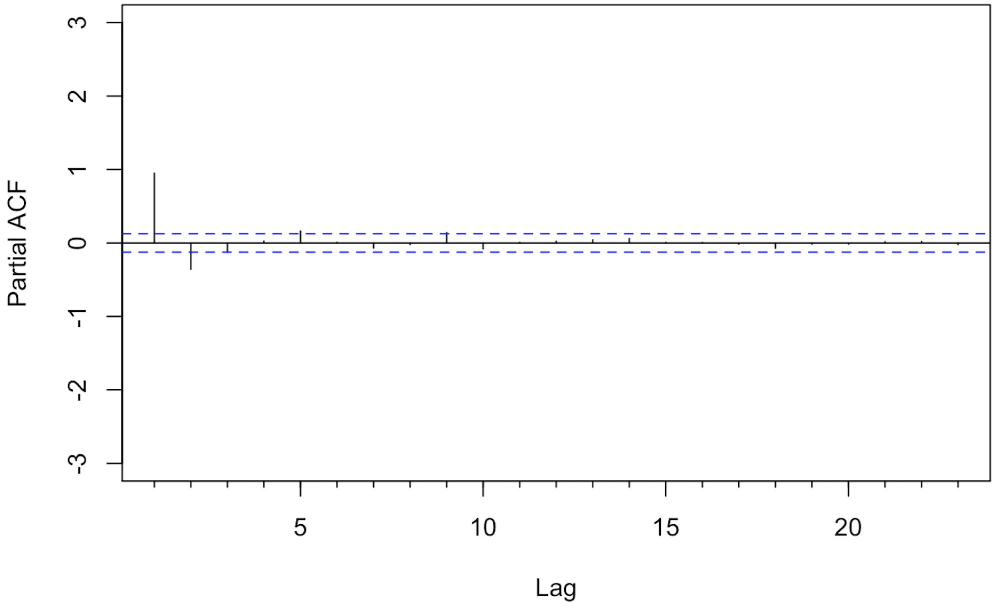

Figure 8: PACF for Unemployment Rate Data

6.2. Issue of Quantile Crossing in DeepProbCP

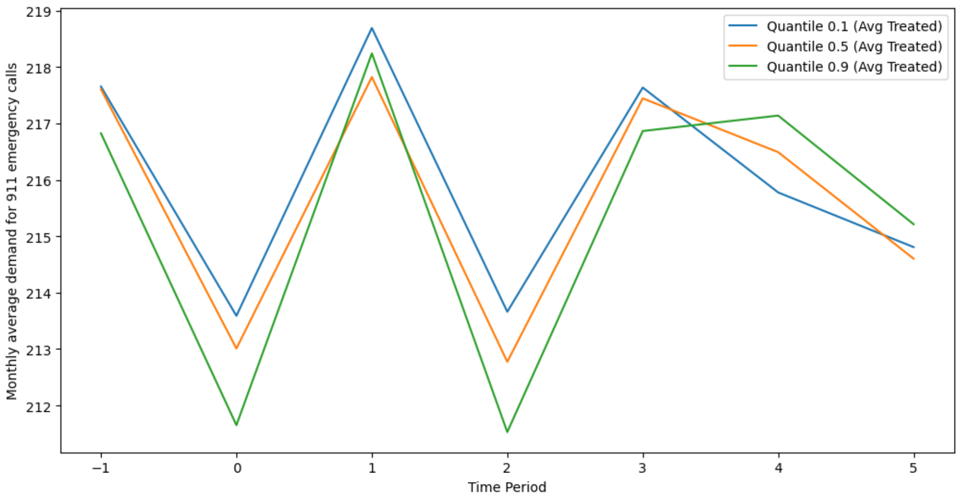

Figure 9: Quantile Crossing in the Treated Units

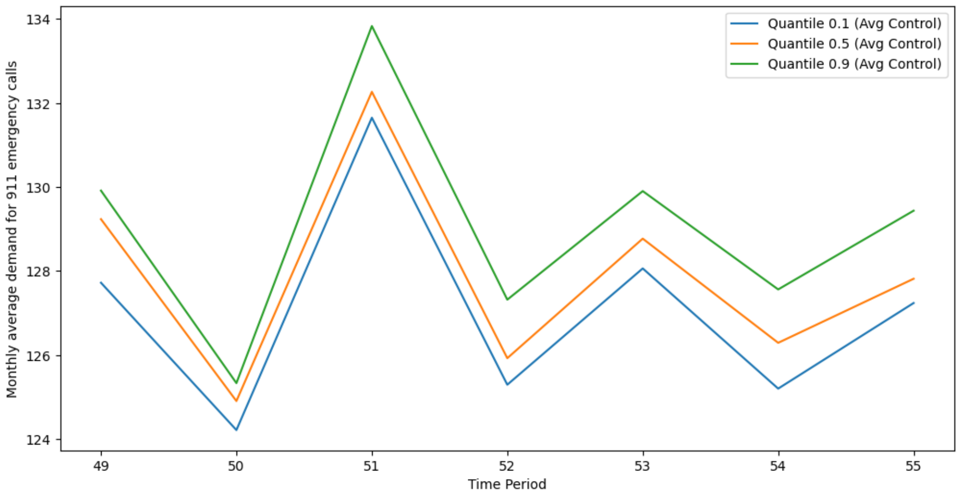

Figure 10: No Quantile Crossing in the Control Units

References

[1] Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, June 2021.

[2] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

[3] Alexis Diamond Alberto Abadie and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

[4] Dionissi Aliprantis. A distinction between causal effects in structural and rubin causal models. Working Papers (Old Series) 1505, Federal Reserve Bank of Cleveland, March 2015.

[5] Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *Journal of Machine Learning Research*, 19(22):1–51, 2018.

[6] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[7] Susan Athey and Guido W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, May 2017.

[8] Kasun Bandara, Christoph Bergmeir, and Hansika Hewamalage. LSTM-MSNet: Leveraging forecasts on sets of related time series with multiple seasonal patterns. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4):1586–1599, apr 2021.

[9] Kasun Bandara, Christoph Bergmeir, and Slawek Smyl. Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach, 2018.

[10] Henry Brady, David Collier, and Jasjeet Sekhon. The neyman-rubin model of causal inference and estimation via matching methods. *The Oxford Handbook of Political Methodology*, 01 2008.

[11] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 9:247–274, 2015.

[12] David Card and Alan B Krueger. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. Working Paper 4509, National Bureau of Economic Research, October 1993.

[13] Carlos Carvalho, Ricardo Masini, and Marcelo C. Medeiros. Arco: An artificial counterfactual approach for high-dimensional panel time-series data. *Journal of Econometrics*, 207(2):352–380, 2018.

[14] Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O. Arik, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting, 2023.

[15] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls, 2021.

[16] Mike Chirico. Emergency - 911 calls, 2020.

[17] Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, 0(0):00491241221099552, 0.

[18] Michelle Degli Esposti, Thees Spreckelsen, Antonio Gasparrini, Douglas J Wiebe, Carl Bonander, Alexa R Yakubovich, and David K Humphreys. Can synthetic controls improve causal inference in interrupted time series evaluations of public health interventions? *International Journal of Epidemiology*, 49(6):2010–2020, 10 2020.

[19] Nikolay Doudchenko and Guido W. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis, 2017.

[20] Xinze Du, Yingying Fan, Jinchi Lv, Tianshu Sun, and Patrick Vossler. Dimension-free average treatment effect inference with deep neural networks, 2021.

[21] Andrew C. Eggers, Guadalupe Tuñón, and Allan Dafoe. Placebo tests for causal inference. *American Journal of Political Science*, n/a(n/a), 2024.

[22] Willem Waegeman Eyke Hüllermeier. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110:457–506, 3 2021.

[23] Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

[24] Jan Gasthaus, Konstantinos Benidis, Yuyang Wang, Syama Sundar Rangapuram, David Salinas, Valentin Flunkert, and Tim Januschowski. Probabilistic forecasting with spline quantile function rnns. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1901–1910. PMLR, 16–18 Apr 2019.

[25] Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151, 2014.

[26] Google-Research. Tft python module. https://github.com/google-research/google-research/tree/master/tft.

[27] Google-Research. Txmiser python module. https://github.com/google-research/google-research/tree/master/tsmixer.

[28] Priscila Grecov, Kasun Bandara, Christoph Bergmeir, Klaus Ackermann, Samuel Campbell, Deborah Scott, and Dan Lubman. *Causal Inference Using Global Forecasting Models for Counterfactual Prediction*, pages 282–294. 05 2021.

[29] Priscila Grecov, Ankitha Prasanna, Klaus Ackermann, Samuel Campbell, Debbie Scott, Dan Lubman, and Christoph Bergmeir. Probabilistic causal effect estimation with global neural network forecasting models.

IEEE Transactions on Neural Networks and Learning Systems, PP:1–15, 07 2022.

[30] Konstantin Hess, Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bayesian neural controlled differential equations for treatment effect estimation, 2023.

[31] Hansika Hewamalage, Christoph Bergmeir, and Kasun Bandara. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1):388–427, 2021.

[32] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[33] Nick Huntington-Klein. Pearl before economists: The book of why and empirical economics. *Journal of Economic Methodology*, 29(4):326–334, 2022.

[34] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Learning and Intelligent Optimization*, 2011.

[35] Paul Hünermund and Elias Bareinboim. Causal inference and data fusion in econometrics, 2023.

[36] Guido W. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79, December 2020.

[37] David Lewis. Causation. *Journal of Philosophy*, 70(17):556–567, 1973.

[38] David Lewis. Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, 2(4):418–446, 1973.

[39] David Lewis. Causation as influence. *Journal of Philosophy*, 97(4):182–197, 2000.

[40] David K. Lewis. *Counterfactuals.* Malden, Mass.: Blackwell, 1973.

[41] Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting, 2020.

[42] Spyros Makridakis, Evangelos Spiliotis, Vassilios Assimakopoulos, Zhi Chen, Anil Gaba, Ilia Tsetlin, and Robert L. Winkler. The m5 uncertainty competition: Results, findings and conclusions. *International Journal of Forecasting*, 38(4):1365–1385, 2022. Special Issue: M5 competition.

[43] Keith A. Markus. Causal effects and counterfactual conditionals: contrasting rubin, lewis and pearl. *Economics amp; Philosophy*, 37(3):441–461, 2021.

[44] The MathWorks. Simulate var model conditional responses, 1994. MathWorks R2021b.

[45] Wang Miao, Xu Shi, and Eric Tchetgen Tchetgen. A confounding bridge approach for double negative control inference on causal effects, 2020.

[46] Nicolaj Søndergaard Mühlbach and Mikkel Slot Nielsen. Tree-based synthetic control methods: Consequences of moving the us embassy, 2021.

[47] Judea Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2), 2010.

[48] Judea Pearl. In defense of unification (comments on west and koch's review of causality). Technical report, UCLA, Los Angeles, CA, 90095-1596, USA, 2014.

[49] Judea Pearl. Causally colored reflections on leo breiman's "statistical modeling: The two cultures" (2001). Technical report, UCLA, Los Angeles, CA 90024 USA, 2021.

[50] Jamal Senouci. Causalimpact python module. https://github.com/jamalsenouci/causalimpact.

[51] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms, 2017.

[52] Claudia Shi, Dhanya Sridhar, Vishal Misra, and David M. Blei. On the assumptions of synthetic control methods, 2021.

[53] Xu Shi, Kendrick Li, Wang Miao, Mengtong Hu, and Eric Tchetgen Tchetgen. Theory for identification and inference with synthetic controls: A proximal causal inference framework, 2023.

[54] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22:1359–1366, 2011.

[55] Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning, 2020.

[56] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, page 567–573, USA, 2002. American Association for Artificial Intelligence.

[57] Granville Tunnicliffe Wilson. Time series analysis: Forecasting and control,5th edition, by george e. p. box, gwilym m. jenkins, gregory c. reinsel and greta m. ljung, 2015. published by john wiley and sons inc., hoboken, new jersey, pp. 712. isbn: 978-1-118-67502-1. *Journal of Time Series Analysis*, 37:n/a–n/a, 03 2016.

[58] Stefan Wager. Stats 361: Causal inference, 2020.

[59] Naftali Weinberger. Comparing rubin and pearl's causal modeling frameworks: A commentary on markus (2021), 2021.

[60] Sewall Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.

[61] Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76, 2017.

[62] Jakob Zeitler, Athanasios Vlontzos, and Ciaran Mark Gilligan-Lee. Nonparametric identifiability and sensitivity analysis of synthetic control models. In *2nd Conference on Causal Learning and Reasoning*, 2023.

Erklärung zur Urheberschaft

Hiermit erkläre ich, Dingyi Lai, dass ich die vorliegende Arbeit noch nicht für andere Prüfungen eingereicht habe. Ich habe die Arbeit selbständig verfasst. Sämtliche Quellen einschließlich Internetquellen, die ich unverändert oder abgewandelt wiedergegeben habe, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, habe ich als solche kenntlich gemacht.

Ich bin mir darüber bewusst, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.

Declaration of Academic Honesty

I, Dingyi Lai, hereby declare that I have not previously submitted the present work for other examinations. I wrote this work independently. All sources, including sources from the Internet, that I have reproduced in either an unaltered or modified form (particularly sources for texts, graphs, tables and images), have been acknowledged by me as such.

I understand that violations of these principles will result in proceedings regarding deception or attempted deception.

Unterschrift / signature Dingyi Lai

Name / name Dingyi Lai

Ort, Datum / place, date Berlin, 18.03.2024